Exploring graph mining approaches for dynamic heterogeneous networks

Lisa Singh
Georgetown University
Computer Science Department
Washington, DC 20057
singh@cs.georgetown.edu

Abstract

As we become a more 'connected' society, a greater need exists to understand complex network structures. While many in the field of data mining analyze network data, most models of networks are straightforward - focusing on many connections of a single type. In order to better understand relationships between different types of entities and extract meaningful structure from heterogeneous data, data mining algorithms need to be developed for new models of complex graphs. In this extended abstract, we describe some existing graph models and propose directions for more robust models that can serve as a backbone for analysis and mining of heterogeneous network data sets. We also identify possible metrics that attempt to capture the relationship of heterogeneous components of the network and serve to give insight into the topological relationships that exist among different node and edge types. The development of these different metrics will be important for designing meaningful clustering and graph mining algorithms for these data. Finally, we consider dynamic versions of these networks and present issues for the next generation of mining algorithms related to community identification, pruning and large graph approximations, privacy preservation of complex graph data.

1. Motivation

For decades, computer science researchers have focused on data sets containing independent and identically-distributed examples. However, a large amount of data in the corporate and scientific domains contains inter-related entities that are linked together in complex graphs. Example application domains include communication networks, protein interaction networks, social networks, transportation systems, and observational scientific networks. The graphs created in these domains are large, feature rich structures involving many different node types and many different edge types.

As an example, we consider a simple observational sciences data set where researchers monitor a subject for a specified period of time. Example subjects include animals, humans, and planets. Each monitoring period can be viewed as an event consisting of a number of observations. Events include tracking an animal, e.g. monkey, for a 30 minute period, conducting a 30 minute psychological evaluation of a person, and taking a five minute snapshot of the interaction between a planet and its moons. Typically, these data sets tend to contain a large number of observations, e.g. thousands to millions, and features, e.g. hundreds to thousands, for a small number of subjects, e.g. tens to thousands. Even though a graph containing all the observations and subjects may be smaller than one involving the Web, the number of features associated with the links and the nodes is very large.

In this simple example, there are three node types (observer, subject, observation). While one can focus on any single node type, integrating the data can help explore methodological questions, data quality issues, and graph mining research questions. Examples include:

- Are observations conducted by different researchers on the team consistent or are there biases?
- Are the observations reliable? Were there field conditions that impacted the quality of the observation? If so, were the field conditions random?
- How are community structures of prominent subjects, exhibiting given behaviors, changing over time?

Because of the large number of features collected by observational researchers, inductive approaches for targeted data exploration are necessary. Researchers can then attempt to better understand community structure and alliance creation, information transmission or behavior propagation through the community, and synergies between genetic relationships and community substructure. This type of analysis necessitates the need to develop data mining algorithms on graph models developed for complex networks.

2. Heterogeneous network models

We now consider the underlying network data model used for network or graph mining. The majority of graph mining algorithms (see [13] for a survey) are developed for *uni-mode* networks, where each node represents a single object type, e.g. an actor, a webpage, or an observation, and each edge represents a single relationship between two nodes in the network, e.g. friendship, kinship, or co-authorship [14]. With the changing structure and increased complexity of network data, a need exists to develop data mining algorithms that support a more generic network model containing multiple node types (multi-mode), multiple edge types (multi-relation), and multiple descriptive features (multi-feature) associated with each.

In previous work, we introduced the M^*3 model, a general representation for a *multi-mode, multi-relation, multi-feature* social networks [8]. This model was originally developed as the basis for visual mining of social networks containing different types of entities and relationships. It allows researchers to visually transform topological structures based on mathematically sound algebras of modes, features and relationships. Visual graph analytic tools using this model are helping biochemists attempting to understand protein interactions and biologists attempting to understand animal behavior using a feature rich data set, investigate multi-entity relationships in the data. This graph centric perspective of the data has many strengths including a more complete representation of real world networks and a straightforward way to incorporate different types of entities and relationships.

After our experiences with the M^*3 model, we discovered that a number of more complex networks do not contain a base set of actors and need to be force fit into the M^*3 model. Further, some networks contain a more complex set of data objects. Given the array of different data objects being integrated together, a more generic model that allow for different object types as well as entity sets can be the basis for future graph mining algorithms. One possible generic model would still allow for multi-mode, multi-relation, multi-feature data, but would extend the concept of multi-feature to include complex data features, e.g. images, time series snippets, etc. This would then allow for network analysis that considers changes of well-structured and complex features associated with nodes and edges.

There are many applications where the inclusion of the complex data could enhance the results of mining activity. The inclusion of photos or sounds for observational scientists is very important for identification of subjects in the wild. For example, if researchers know that certain monkeys are seen regularly together and suddenly the group composition changes, photos can be used to see if bites or injuries on a monkey may have caused misclassification or if the community structural change actually occurred. Using network connectivity in conjunction with image data can help increase the quality of observational results and even help correct errors related to potential duplication.

Another important extension is the specific transformation that optimizes the generic network model for dynamic or time-varying analysis. While investigating changes in community structure over time is a current research area [12, 3, 2], the networks being used for these analyses contain a single node and single edge type. We want to expand the analyses to incorporate complex graph models. Questions of interest include:

- How stable are these complex graph structures over time?
- Which modes or relationships change most frequently?
- What is the topological difference between a multi-featured community at time t_1 and at time t_2?

Community extraction algorithms need to extend current pattern mining approaches to consider multi-featured, heterogeneous networks. A thread of work exists in mining hidden communities in heterogeneous social networks containing multiple relationship types [5]. We have begun looking at dynamic bi-mode affiliation networks to better understand changing community dynamics [7]. Here we visually track the group structure of a pair of actors to better understand the changing group dynamic over time. However, this is just a first step. Integrating longitudinal dynamics with complex, heterogeneous networks is an outstanding challenge. One large reason is the sheer volume of data generated. We will address the challenge of data size in Section 5.

3. Developing metrics for understanding complex structures

In order to mine heterogeneous networks, we need to develop measures that can be used to better understand the specific network under consideration. While the uni-mode measurements will remain important, extensions for numerous modes and numerous relationships are needed. We have begun to develop measures for multi-mode density and multi-mode k-awayness (or hop expansion) [9]. For example, the multi-mode density is zero if all the edges are in a single mode. As the number of edges connecting nodes in different modes increases, the modal-density measure also increases.

Building on this work, we suggest developing measures that take varying edge types, nodes types, and feature distributions into consideration. Possible examples include

multi-edge path length (the number of edges traversed between two nodes in a multi-relational path), strength of connections (frequency and duration of different types of interactions), transmission rate (paths of feature value expansion across different feature values, e.g. foraging behaviors), and network turnover (a longitudinal measure that captures affiliation changes over time).

With these topological metrics, we can use structure to understand the growth distribution of temporal generic networks. We can analyze how growth rates of one mode affect the growth rates of others and consider how different attribute features affect the structural properties of the network? While physicists have been studying the dynamics of network formation and growth [10, 4], only the simplest of models are understood. Those findings need to be extended to more complex structures and specialized metrics for measuring individual, local community and global network statistics, will be vital for understanding the strengths and weaknesses of the generic model.

4. Directions for extending graph mining algorithms

Unfortunately, since graph mining algorithms focus on uni-mode models and traditional mining algorithms do not consider network topology during the process, a need exists to develop new algorithms that extend traditional clustering and community extraction algorithms by incorporating different modes of data in a meaningful way. Issues that need addressing include:

- What metrics are meaningful when clustering different entity types?
- How does mode topology influence different clustering algorithms? Can this be easily quantified?
- Are there certain data structures that help us quickly identify similar nodes based on topological location?
- Are relational operators such as union, minus, and intersect useful for adjusting network structure prior to mining?
- What is the impact of variance in mode size and can we determine which modes to ignore based on connectivity structure or feature distribution?

5. Pruning complex networks

As previously mentioned, an additional challenge of this approach is the volume of data that needs to be analyzed when multiple relations are combined. Developing methods for identifying 'important' subgraphs from which to identify community substructure makes the problem more tractable. This step is necessary if we are interested in classifying network objects based on topological substructure and object features. Several proposed methods for classification of network objects consider the link structure of the network [6].

When considering link based approaches for heterogeneous networks, the logical relationship between objects and the probabilistic dependencies between attributes may cause a huge search space for subgraph mining. Previous work [11] shows that predictive accuracy can be maintained on affiliation network (two-mode network) objects if instead of random pruning, the network under consideration is pruned based on attribute values and/or structural properties like degree and betweenness. Also, pruning the data prior to analysis can remove noisy components of the data. By removing some of the less relevant components of the data, we can improve predictive accuracy of classifiers and extract smaller, more meaningful clusters and graph substructures from the data.

We now need to reinvestigate whether or not these findings still apply for more complex networks? Is structure less of a predictive indicator for multi-mode networks? Given that no one pruning approach will work across data sets, we propose selecting a pruning approach prior to subgraph extraction and classification based on local structural graph invariants (hop expansion, clique structure, clustering coefficient, etc.) and node specific feature measures (behaviors, gender, lineage, etc.). We can then compare the structural similarities and the predictive accuracies of these pruned networks to the full network to better understand the strengths and weaknesses of the different pruning strategies on generic networks.

6. Complex social networks and privacy

While privacy preservation of data mining approaches has been an important topic for a number of years, privacy of social network data is a relatively new area of interest. Typically when we consider social network data, we view it as data that is available to the public. However, many social networks are now being automatically extracted from private data sources. Examples include social networks derived from corporate email servers, customer referral databases, personal medical records, and disease population databases.

In this area, what constitutes a privacy breech is still an open question. However, previous research have shown that anonymization alone may not be sufficient for hiding identity information on certain real world data sets [1]. To date the research conducted in this arena has been on uni-mode networks. If we consider complex networks, identification of individuals in the network becomes easier. How do we

combat this? To what degree is network topology a factor compared to node and edge features? Are relationships between nodes more apparent when local neighborhoods have certain topological structures? How can we use the topological structure of complex networks to measure the level of anonymity in the network? Finally, what measures are reasonable for quantifying various levels of the topology? To study some of the behaviors associated with social networks, how accurate do the network measures need to be for data mining applications, e.g. clustering, community discovery, prominent node identification, etc.? While we anticipate many of these topics will be explored soon for uni-mode networks, a far-reaching goal is to consider privacy preservation in the context of dynamic, heterogeneous networks.

7. Final thoughts

While we are still investigating ways to analyze simple networks with a single node type and a single edge type, the complexity of today's network data forces us to begin thinking of ways to handle and analyze more heterogeneous data. In order to mine the data, we need to develop robust models that capture the interconnected nature of the data, while allowing for the inclusion of complex features and time varying attributes. In this abstract, we proposed some directions and identified a set of issues associated with mining data using a generic model for dynamic, heterogeneous networks. We describe some of the issues in the context of an observational scientific data set and alluded to other complex network data sets. Finally, we pose a number of questions that need to be considered when working with dynamic, heterogeneous networks for different data mining applications.

References

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 181–190, New York, NY, USA, 2007. ACM Press.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, New York, NY, USA, 2006. ACM Press.

[3] T. Y. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 523–528, New York, NY, USA, 2006. ACM Press.

[4] M. Boguna and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical review*, E 66(4), 2002.

[5] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Mining hidden community in heterogeneous social networks. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 58–65, New York, NY, USA, 2005. ACM Press.

[6] L. Getoor. Link-based classification. In S. Bandyopadhyay, U. Maulik, L. Holder, and D. Cook, editors, *Advanced Methods for Knowledge Discovery from Complex Data*. Springer, 2005.

[7] L. S. Hyunmo Kang, Lise Getoor. C-group, a visual analytic tool for pairwise analysis of dynamic group membership. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, 2007.

[8] L. G. Lisa Singh, Mitchell Beard and M. Blake. Visual mining of multi-modal social networks at different abstraction levels. In *IEEE Conference on Information Visualization*, 2007.

[9] G. Nelson. Using neighborhood measures for visual mining of multi-modal graphs, 2007.

[10] M. J. Newman. Clustering and preferential attachment in growing networks. *Physics Review*, 64, 2001.

[11] L. Singh, L. Getoor, and L. Licamele. Pruning social networks using structural properties and descriptive attributes. In *IEEE International Conference on Data Mining*, pages 773–776, Washington, DC, USA, 2005. IEEE Computer Society.

[12] C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 717–726, New York, NY, USA, 2007. ACM Press.

[13] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Explor. Newsl.*, 5(1):59–68, 2003.

[14] S. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, Cambridge, 1994.